

Tsvetana Dimitrova

Institute for Bulgarian Language, Bulgarian Academy of Sciences
Sofia, Bulgaria

ISSUES IN THE LINGUISTIC ANNOTATION OF DIACHRONIC CORPORA

(Summary)

The paper discusses the impact of graphic variants, wrongly spelled constructions and some ambiguous word forms, found in the Old Bulgarian texts, on the annotation practice in diachronic language corpora. Markup choices are illustrated by examples from two corpora, containing Codex Marianus texts. The manuscript exemplifies most problems that are common in the annotation of diachronic texts.

Keywords: diachronic language corpora, linguistic annotation, annotation layers, tokenization, graphic variants, scripta continua, errors, disambiguation.

✉ Tsvetana Dimitrova
tsvetana.dimitrova@gmail.com

Published: 31 March 2010